The Two-Sample Linear Regression Model with Interval-Censored Covariates

David Pacini

University of Bristol*†

June 2018

Abstract

There are surveys that gather precise information on an outcome of interest, but measure continuous covariates by a discrete number of intervals, in which case the covariates are *interval-censored*. For applications with a second independent dataset precisely measuring the covariates, but not the outcome, this paper introduces a semiparametrically efficient estimator for the coefficients in a linear regression model. The second sample serves to establish point identification. An empirical application investigating the relationship between income and body mass index illustrates the use of the estimator.

KEYWORDS: interval-censoring, sample combination, semiparametric efficiency.

JEL codes: C21, C24.

*Address: Department of Economics, University of Bristol, 8 Woodland Road, BS8 1TN, Bristol UK. Email: david.pacini@bristol.ac.uk. Tel. +44 (0) 1173317937.

†Acknowledgments. The manuscript has benefited from the comments by Gregory Jolivet and participants at the Bristol Econometric Study Group '17, the Econometric Society European Meeting '17, and the Econometric Workshop at Toulouse School of Economics.

1. Introduction

Continuous covariates are measured in several surveys by a discrete number of intervals, e.g., income between \$0 and \$9,999, \$10,000 and \$19,999, etc. Examples of this type of imprecise measurement, so-called interval-censoring, include income in the Health Survey for England (HSE), the General Social Survey, the American Time Use Survey, the National Health and Nutrition Examination Survey, the National Adult Tobacco Survey, the UK National Travel Survey, and wages in the UK Workplace Employee Relations Survey. On one hand, interval-censoring has advantages because it reduces nonresponse and/or protects privacy. On the other hand, interval-censoring is problematic because, by distorting the genuine dispersion of the covariates, it makes standard inference procedures inaccurate and imprecise, i.e., it entails a loss of identification and power when estimating and testing.[1]

This trade-off has prompted the development of nonstandard one- and two-sample inference procedures. One-sample procedures, in turn, have developed into two types. In one case, the procedure is based on a point-identifying approach parameterizing the distribution of the covariates.[2] In the other case, the procedure is based on a set-identifying approach that does not rely on parametric assumptions.[3] While the former offers precision gains over the latter, under an incorrect parametrization, these gains may not be realized or, even worse, may be transformed into an accuracy loss. Two-sample procedures have resulted from applications with a second dataset having continuous measurements of the covariates.[4] For instance, the UK Family Resources Survey (FRS) has continuous measurements for income, which is interval-censored in the HSE. So far, two-sample procedures have only followed the

[1]For the advantages, see, e.g., Juster and Smith (1997). For the disadvantages, see, e.g., Hsiao (1983) and Rigobon and Stoker (2009).

[2]See, e.g., Hsiao (1983) for a linear regression model.

[3]See, e.g., Beresteanu, Molchanov and Molinari (2011) and Bontemps, Magnac and Maurin (2012) for linear projections; Magnac and Maurin (2008) and Bhattacharya and Lee (2018) for binary response models; Manski and Tamer (2002) and Cerquera, Laisney and Ullrich (2015) for monotone regression models.

[4]See, e.g., Pollmann (2015), Asher, Novosad and Rafkin (2018) and, for the income variable in the National Travel Survey, see Lepanjuuri, Cornick, Byron, Templeton and Hurn (2016).

set-identifying approach. Little is known about how two samples can realize the precision gains offered by the point-identifying approach without parameterizing the distribution of the interval-censored covariates.

This paper introduces an estimator, called the two-step two-sample augmented generalized instrumental variable (2S-AGIV) estimator, drawing on the comparative advantages of the point- and set-identifying approaches. The paper has three results. The first result states that, when there is a second independent sample with continuous measurements of the covariates, the linear regression model with interval-censored covariates in the first sample point identifies the coefficients of interest. The model implies an identifying moment restriction using indicator variables for the intervals as instrumental variables observed in both samples. Neither parametric, support, nor monotonicity restrictions on the interval-censored covariates are needed to obtain this result. The second result shows that the existing two-sample instrumental variable estimators, including the two-stages least squares (2SLS, Klevmarken, 1982) and two-sample instrumental variable (2S-GIV, Ridder and Moffitt, 2007) estimators, are consistent and asymptotically normal; however, they are not semiparametrically efficient. The 2SLS estimator is equivalent to imputing in the censored sample the truncated mean of the covariate of interest within the interval calculated from the uncensored sample. The 2S-GIV estimator is equivalent to a weighted least squares estimator on the truncated mean outcome and covariate of interest within the interval. The paper shows that the 2S-AGIV estimator is consistent, asymptotically normal, and semiparametrically efficient, which is the third result. This last property means that, in large samples, the 2S-AGIV estimator can realize in the best possible way the precision gains offered by the point-identifying approach without parameterizing the distribution of the covariates. A simulation study bears out these theoretical properties of the 2S-AGIV estimator.

An empirical exercise using data from the HSE and the FRS illustrates and supports the use of the 2S-AGIV estimator. The exercise tests the Unearned Income Effect (UIE)

hypothesis, which postulates an inverted U-shaped relationship between income and body mass index (see Lakdawalla and Philipson, 2009). This hypothesis is relevant, for instance, in assessing the effects of redistributive polices on obesity. In the HSE, body mass index is precisely measured but income, the covariate of interest, is interval-censored. The FRS has continuous measurements for income, but not for body mass index. When applied to the subsample of female adults, the 2S-AGIV strongly rejects the UIE hypothesis. The rejection of the UIE is not a weakness of the 2S-AGIV estimator. Rather, it highlights one relevant capability of this estimator, namely, the possibility to alert the applied researcher to the fragility of the conclusions reached from the UIE. The strong rejection of the UIE for the subsample of females is not delivered by existing one-sample procedures.

The rest of this paper is organized as follows. We next review the related literature. Section 2 sets up the two-sample model with interval-censored covariates. It shows that this model point identifies the coefficients of interest. Section 3 describes the 2S-AGIV estimator. It establishes that this estimator is semiparametrically efficient. Section 4 illustrates the use of the 2S-AGIV estimator. Section 5 presents the results from the simulation study. Section 6 concludes and outlines extensions. Appendix A collects the proofs of the propositions. Appendix B describes the two-sample misspecification tests for the validity of the linear regression model. There are two online appendices. Appendix C presents the proofs of the lemmas. Appendix D presents further the results from the simulation study.

The study of interval-censored covariates can be traced back to Hsiao (1983), with subsequent contributions by Manski and Tamer (2002, see also Cerquera et al., 2015), Magnac and Maurin (2008), Rigobon and Stoker (2009), Beresteanu et al. (2011), Pollmann (2015), Kaido (2017), and Asher et al. (2018). We are the first to address the problems of identification and efficient estimation in the two-sample linear regression model.[5] Hsiao (1983)

[5]As the authors indicate, the two-sample model in Devereux and Tripathi (2009, see footnote 8) does not cover interval-censoring.

and Rigobon and Stoker (2009) document the pitfalls of using naive methods to deal with interval-censored covariates. Hsiao (1983) proposes a strategy based on parameterizing the distribution of the covariates. Such a strategy delivers point-identification and enables maximum likelihood estimation. Manski and Tamer (2002) pionnered the alternative strategy of maintaining no distributional assumption, at the cost of losing point-identification, and performing a worst-case analysis (see also Magnac and Maurin, 2008; Beresteanu, Molchanov, and Molinari, 2011; Pollmann, 2015; Kaido, 2017; Asher et al., 2018). When a second independent sample with precise measurements on the covariates is available, we show that, in the linear regression model, one can maintain no distributional assumption and still achieve point-identification.

The 2S-AGIV estimator is related to the literature on two-sample instrumental variable estimators.[6] The 2S-AGIV can be seen as an extension of the 2SLS and 2S-GIV estimators. Neither the 2SLS nor the 2S-GIV estimator have so far been applied to deal with interval-censored covariates. Both the 2SLS and 2S-GIV are consistent and asymptotically normal. This paper shows that, unlike the 2SLS and the 2S-GIV estimators, the 2S-AGIV estimator is semiparametrically efficient. We use ideas from the missing covariate literature to construct the 2S-AGIV estimator. Since the missing covariate literature so far has not dealt with the interval-censored covariate problem, this new use of existing ideas is a novelty of the paper.[7] For the missing covariate problem, it is well-known, see e.g., Graham (2011), that a semiparametrically efficient estimator can be constructed from the so-called augmented moment function. The 2S-AGIV estimator is developed after constructing the augmented moment function for the two-sample interval-censored covariate problem. The augmented moment function uses a transformation of the interval-censored covariate as instrumental

[6]See Ridder and Moffitt (2007) for a review and Inoue and Solon (2010), Pacini and Windmeijer (2016) and Choi, Gu and Shen (2017) for more recent developments.

[7]See, e.g., Robins, Rotnitzky and Zhao (1994), Chen, Hong and Tarozzi (2008), Graham (2011), Graham, Pinto and Egel (2012), Dardanoni, De Luca, Modica and Peracchi (2015), and Chaudhuri and Guilkey (2016).

variables because in the censored sample, only the interval-censored version of the covariate is observed.

2. The Linear Regression Model and the Two Samples

The aim is to construct a semiparametrically efficient estimator for the unknown coefficient β_o in the two-sample linear regression model with an interval-censored covariate, to be defined below. Let S_C and S_U denote two lists of observational units. They have, respectively, n_C and n_U units - the sample's sizes. Let $S = S_C \cup S_U$ and $n = n_C + n_U$ denote the union and the total number of observational units, respectively. Let $1(\cdot)$ denote a function taking the value of one when the condition in parentheses is valid, and zero otherwise. For any column random vectors r and s, let $E(r)$, $E(r|s)$, $V(r)$ and $V(r|s)$ denote their unconditional and conditional expectation and variance, respectively. The following assumptions serve to define β_o. They will be used to approximate the behavior of the estimators of β_o.

Assumption 1 (Linear Regression). (a) $y = x\beta_o + u$, (b) $E(u|x) = 0$, and (c) $0 < E(x^4) < 0$, $0 < E(u^4) < \infty$, where y and x are the outcome and the continuous scalar covariate of interest, respectively, and u is a disturbance term.

Assumption 2 (Interval-Censored Covariate). Let $B > 0$ be an integer less than the number of points in the support of x. Define the partition $L_1 < U_1 \leq L_2 ... < L_b < U_b \leq L_{b+1} .. \leq L_{B+1} < U_{B+1}$. There are known surjective functions g_L and g_U, with support $\{L_1, .., L_b, .., L_{B+1}\}$ and $\{U_1, .., U_b, .., U_{B+1}\}$, respectively, such that, for $\underline{x} = g_L(x)$ and $\overline{x} = g_U(x)$, x belongs to the interval $[\underline{x}, \overline{x}]$ with probability one.

Assumption 3 (Interval-Censored Sample). There is an interval-censored sample $\{y_i, \underline{x}_i, \overline{x}_i\}_{i \in S_C}$ with n_C independent and identically distributed (i.i.d.) replications of $(y, \underline{x}, \overline{x})$. Define $Y_C := \{y_i\}_{i \in S_C}$.

Assumption 4 (Uncensored Sample). There is an i.i.d. sample $X_U := \{x_j\}_{j \in S_U}$ with n_U replications of x.

Assumption 5 (Interval-Censored and Uncensored Samples are Independent). For $l = 1, .., n$, define $d_l = 1$ if the l-th drawn corresponds to the censored sample S_C, and zero otherwise. (a) d is independent of (y, x) and $0 < \kappa_o := E(d) < 1$; (b) $S_C \cap S_U = \emptyset$; and (c) S_C and S_U come from the same population.

Assumption 6 (Full Rank). Let $w := \big(1(\underline{x} = L_1)1(\overline{x} = U_1), .., 1(\underline{x} = L_B)1(\overline{x} = U_B)\big)'$ be a $B \times 1$ vector of dummy variables indicating the interval to which a realization of x belongs. $E(ww')$ and $E(wx)$ have full rank.

Joint data on (y, x) are not available in this model. The model places no restriction on either the distribution or the support of x, except for ruling out a binary x. Several remarks are in order.

Remarks on Assumption 1. To simplify the exposition, x is a scalar, although this restriction can be relaxed. In the case of having more than one interval-censored covariate, the results below carry through after stacking in w_i the list of indicator variables for the intervals of each interval-censored covariate. In the case of having control covariates observed in both samples, y_i and x_i would be residuals from regressing the original outcome and covariate of interest on the chosen control covariates.

Remarks on Assumption 2. First, since B is smaller than the number of elements in the support of x, $g_L(\cdot)$ and $g_U(\cdot)$ are non-injective. Hence, even if $g_L(\cdot)$ and $g_U(\cdot)$ are known, one cannot get x from knowing x_L and x_U. Second, Assumption 2 excludes incorrect interval reporting due to bracketing effects, which may be present in some surveys.[8]

Remarks on Assumption 4. The outcome is not observed in the uncensored sample, that is, there is no complete case subsample.[9] From the uncensored sample, one can calculate the bounds $\underline{x}_j := g_L(x_j)$ and $\overline{x}_j := g_U(x_j)$ for $j \in S_U$. Since w only depends on $(\underline{x}, \overline{x})$, it is

[8] See, e.g., Winter (2002).

[9] The absence of a complete subsample is one of the differences between the two-sample linear regression model and the linear models studied in the missing data literature, c.f., Dardanoni et al. (2015) and Chaudhuri and Guilkey (2016).

observed in both samples.

Remark on Assumption 5. The independence between d and (y, x) does not imply -nor is it implied- by the conditional independence restriction in, for example, Chen et al. (2008, Assumption 2). Since $S_C \cap S_U = \emptyset$, the uncensored sample is not a subsample of the interval-censored sample.

Remark on Assumption 6. To interpret the full rank assumption, from the definition of w, notice that $E(ww')$ is a $B \times B$ diagonal matrix with a characteristic b-th element equal to the proportion of observational units in the interval $[L_b, U_b]$. The full rank restriction on $E(ww')$ thus excludes populations with no observational units in a given interval.[10] The b-th element in $E(wx)$ is the truncated mean of x in the interval b divided by the proportion of observational units in that interval, i.e., $E(w_b x) = E(x|L_b \leq x \leq U_b)/E(w_b)$. The full rank restriction on $E(wx)$ rules out having fewer intervals than coefficients in β_o. Since β_o is a scalar, Assumption 6 holds only if $B \geq 1$.

Assumption 6 differentiates the missing and the interval-censored covariate problems. In the former problem, since one can represent the support of x as the interval with endpoints L_1 and U_1, $B = 0$. The transformation w of x cannot play any role because the full rank assumption cannot be satisfied with $B = 0$. By contrast, the interval-censored covariate problem opens the possibility for w to play a role when making inferences about β_o.

2.1 Point Identification

When only the interval-censored sample is available, β_o only can be set identified.[11] The next result shows that, by contrast, the two-sample model point identifies β_o.

[10] As pointed out by a referee, one may wonder whether there are instruments other than w_i observed in both samples, such as the midband of the interval, which can be used to increase the precision when estimating β_o. The midband, like any other point in the interval, is a linear combination of w_i. Hence, the use, on top of w_i, of the midband will not offer precision gains.

[11] The population regressions of y on $\underline{x}$ and y on $\overline{x}$ do not, in general, lead to the same value of the coefficient, which shows that we cannot identify β_o from the joint distribution of $(y, \underline{x}, \overline{x})$.

Proposition 1 (Point Identification). *Assumptions 1 to 6 deliver the point identification of β_o. Since w is a transformation of x, and w is binary,*

$$E(y - x\beta_o|w) = 0 \text{ if and only if } E\left[\frac{d}{\kappa_o}wy - \frac{(1-d)}{(1-\kappa_o)}wx\beta_o\right] = 0. \tag{WMR}$$

The point-identifying Weighted Moment Restriction (WMR) combines the samples. Proposition 1 formalizes the intuition that the group means $E(y_i|w_i)$ and $E(x_i|w_i)$ point identifies β_o. Since u is mean-independent of w, one has $E(y|w) = E(x|w)\beta_o$, where the censored sample identifies $E(y|w)$, and the uncensored sample identifies $E(x|w)$. w_i plays the role of a list of instrumental variables observed in both samples. To consistently estimate β_o, is sufficient to calculate the ordinary least squares estimator on the sample analog of $E(y_i|w_i)$ and $E(x_i|w_i)$. The discussion below shows that this strategy does not combine the information in the two samples efficiently.

3. Semiparametric Efficient Estimation

Since WMR is a two-sample linear moment restriction, with instruments observed in both samples that may overidentify the parameter of interest, one can estimate β_o by either the 2SLS or the 2S-GIV estimators. We next describe these two estimators.

3.1 The 2SLS and 2S-GIV Estimators

Applying the formula for the 2SLS estimator to WMR yields:

$$\hat{\beta}_{2sls} := (X_U'W_U\hat{\Omega}_U\hat{\Omega}_C^{-1}\hat{\Omega}_UW_U'X_U)^{-1}X_U'W_U\hat{\Omega}_U\left(\frac{n_U}{n_C}W_C'Y_C\right), \tag{2SLS}$$

with $W_U := \{w_j\}_{j\in S_U}$, $W_C := \{w_i\}_{i\in S_C}$, $\hat{\Omega}_U := (W_U'W_U/n_U)^{-1}$, and $\hat{\Omega}_C := (W_C'W_C/n_C)^{-1}$. The 2SLS estimator is equivalent to a linear two-stage imputation or regression calibration

method.[12] The first stage consists of imputing, in the censored sample, the truncated mean of the covariate calculated from the uncensored sample, that is, $\hat{X}_C = W_C(W_U'W_U)^{-1}W_U'X_U$. The second stage consists of calculating the ordinary least squares estimator with the imputed data, that is, $\hat{\beta}_{2sls} = (\hat{X}_C'\hat{X}_C)^{-1}\hat{X}_C'Y_C$.

Applying the formula of the two-sample estimator in Ridder and Moffitt (2007, Formula 86) yields the family of GIV estimators:

$$\hat{\beta}_{\hat{\Omega}} := (X_U'W_U\hat{\Omega}W_U'X_U)^{-1}X_U'W_U\hat{\Omega}\left(\frac{n_U}{n_C}W_C'Y_C\right), \qquad \text{(GIV Family)}$$

where $\hat{\Omega}$ is any $B \times B$ positive definite matrix.

To describe the estimator in the GIV family with the smallest asymptotic variance, the following result is needed.

Lemma 1 (GIV Family - Consistency and Asymptotic Normality). Let Assumptions 1 to 6 hold. Furthermore, assume that the weighting matrix $\hat{\Omega}$ converges in probability to a positive definite matrix Ω_o. Then, $n^{1/2}(\hat{\beta}_{\hat{\Omega}} - \beta_o)$ converges in distribution to a zero-mean normal random vector, denoted as $n^{1/2}(\hat{\beta}_{\hat{\Omega}} - \beta_o) \rightsquigarrow \mathcal{N}\big(0, avar(\hat{\beta}_{\hat{\Omega}})\big)$, with variance

$$avar(\hat{\beta}_{\hat{\Omega}}) := [E(xw)\Omega_o E(wx')]^{-1}[E(xw)\Omega_o\Sigma_o\Omega_o E(wx')][E(xw)\Omega_o E(wx')]^{-1}, \qquad (1)$$

where $\Sigma_o := V(wy)/\kappa_o + V(wx'\beta_o)/(1-\kappa_o)$.

[12] For a book treatment of the regression calibration method, see Carroll, Ruppert, Stefanski and Crainiceanu (2006, Chapter 4). The efficiency properties of the two-sample 2SLS estimator, and similar regression imputation procedures, have so far not been explored.

For any given β, define

$$\hat{\Sigma}_\beta := \frac{n}{n_C^2}\sum_{i\in S_C} w_i y_i^2 w_i' - \frac{n}{n_C^3}(W_C'Y_C)(W_C'Y_C)' + \frac{n}{n_U^2}\sum_{j\in S_U} w_j (x_j'\beta)^2 w_j' - \frac{n}{n_U^3}(W_U'X_U\beta)(W_U'X_U\beta)'.$$

The matrix Ω_o minimizing $avar(\hat{\beta}_{\hat{\Omega}})$ is the inverse of Σ_o, which can be consistently estimated by the inverse of $\hat{\Sigma}_{\hat{\beta}_{2sls}}$. The GIV estimator with the weighting matrix $\hat{\Sigma}^{-1}_{\hat{\beta}_{2sls}}$ is referred to as the Two-Step GIV (2S-GIV) estimator and is denoted as $\hat{\beta}_{giv}$. Its asymptotic variance is[13]

$$avar\big(\hat{\beta}_{giv}\big) := [E(xw')\Sigma_o^{-1}E(wx)]^{-1}. \tag{2}$$

The 2S-GIV estimator can be interpreted as a weighted least squares estimator based on regressing the sample analog of $E(y|w)$ on the sample analog $E(x|w)$ calculated from the censored and uncensored samples, respectively.

The estimators in the GIV family use only one weighting matrix. The 2SLS estimator does not belong to the GIV family because it uses two weighting matrices $\hat{\Omega}_C$ and $\hat{\Omega}_U$, which, in general, do not coincide. This contrasts with the one-sample 2SLS estimator, which belongs to the family of one-sample GIV estimators. However, since $\hat{\Omega}_C$ and $\hat{\Omega}_U$ converge in probability to the same limit $E(ww')^{-1}$, $\hat{\beta}_{2sls}$ has asymptotic variance $avar(\hat{\beta}_{\hat{\Omega}})$, as defined in Lemma 1 with Ω_o replaced by $E(ww')^{-1}$. The 2SLS is then, in general, less precise than the 2S-GIV. The discussion below suggests that neither the 2SLS nor the 2S-GIV estimator are asymptotically efficient.[14]

3.2 The Semiparametric Efficiency Bound

The next proposition establishes the semiparametric efficiency bound for regular estimators of β_o. This bound serves to qualify and quantify the efficiency loss incurred by the 2SLS

[13]The asymptotic variances in (1) and (2) are special cases of Ridder and Moffitt (2007, Formula (179))

[14]For more on the relationship between the 2SLS and two-sample GMM estimators, see Pacini and Windmeijer (2016).

and 2S-GIV estimators and look, if necessary, for more efficient alternatives.

Lemma 2 (Efficiency Bound). Let Assumptions 1 to 6 hold. Define the $B \times B$ matrix $\Psi_o := V[wE(x\beta_o|w)]/[(1-\kappa_o)\kappa_o]$. The maximal asymptotic precision with which β_o may be regularly estimated is given by

$$\mathcal{I}_o^{-1} := \left[E(xw')(\Sigma_o - \Psi_o)^{-1}E(wx')\right]^{-1}, \qquad \text{(Efficiency Bound)}$$

where Σ_o is defined in Lemma 1.

$\mathcal{I}_o^{-1}$ is a special case of the efficiency bound in Graham et al. (2016), which so far has not been related either to the interval-censored covariate problem, the 2SLS or the 2S-GIV estimator.

An inspection of $avar\left(\hat{\beta}_{giv}\right)$ and $\mathcal{I}_o^{-1}$ reveals the following result:

Corollary (Inefficient Estimators). In general, the 2SLS and the 2S-GIV estimators do not attain the semiparametric efficiency bound.

The asymptotic inefficiency of the 2S-GIV estimator depends on: the value of the coefficient of interest β_o, the proportion of censored observations κ_o, and the distribution of (x, w).[15] The inefficiency increases with the absolute value of β_o. When $\beta_o = 0$, the 2S-GIV has no efficiency loss. When κ_o differs from one half, i.e., an increase in $V(d)$, the inefficiency of the 2S-GIV increases. The simulation study in Section 5 quantifies the inefficiency for specific distributions of (x, w).

[15] Inoue and Solon (2010) establish that the just-identifying 2SLS estimator is more precise than the just-identifying two-sample instrumental variable estimator, while Graham et al. (2011) establish that the latter estimator is not efficient. These two results are not sufficient to establish the inefficiency of the 2SLS and the 2S-GIV estimators.

3.3 A Semiparametrically Efficient Estimator

To construct a semiparametrically efficient estimator, following ideas from the missing covariate literature, e.g., Graham (2011), the strategy is to derive the augmented weighted moment restriction of the interval-censored covariate problem. As shown in Proposition 1, the model implies the moment restriction WMR. On top of WMR, the independence of d and w, from Assumption 5(a), implies the Augmenting Moment Restriction (AMR)

$$E[(d-\kappa_o)g(w)]=0, \tag{AMR}$$

for any function $g(\cdot)$ with $E[|g(w)|]$ finite. One strategy to exploit the information in WMR and AMR is to reduce the sampling variation in WMR by subtracting AMR after having chosen $g(\cdot)$, such that the moment function in this difference has a variance equal to the inverse of the term in the middle of the efficiency bound in Lemma 2. This strategy suggests the Augmented Weighted Moment Restriction (AWMR)

$$E\left[\frac{wyd}{\kappa_o}-\frac{wx\beta_o(1-d)}{(1-\kappa_o)}-\frac{(d-\kappa_o)w\mu_o(w)}{(1-\kappa_o)\kappa_o}\right]=0, \tag{AWMR}$$

where $\mu_o(w):=E(y|w)$. The next lemma verifies that the variance of the moment function in AWMR is the expected one.

Lemma 3 (Variance - AWMR). Let Assumptions 1 to 6 hold. The variance of

$$\frac{wyd}{\kappa_o}-\frac{wx\beta_o(1-d)}{(1-\kappa_o)}-\frac{(d-\kappa_o)w\mu_o(w)}{(1-\kappa_o)\kappa_o}$$

is $\Upsilon_o:=\Sigma_o-\Psi_o$*, where* Σ_o *and* Ψ_o *are defined in Lemmas 1 and 2, respectively.*

From comparing the moment WMR delivering point-identification with the moment AWMR

delivering a semiparametrically efficient estimator, one can conclude that constructing a semiparametrically efficient estimator requires more elaboration than simply regressing the sample analog of $E[y|x \in (\underline{x}, \overline{x})]$ on $E[x|x \in (\underline{x}, \overline{x})]$.

Lemma 3 suggests that, if Υ_o and μ_o were known, a one-step semiparametrically efficient estimator of β_o would be

$$\hat{\beta}_{\Upsilon_o,\mu_o} := \left(X_U' W_U \Upsilon_o^{-1} W_U' X_U\right)^{-1} X_U' W_U \Upsilon_o^{-1} \left(\frac{n_U}{n_C} W_C' Y_C - \frac{n_U}{n} W' A_o\right),$$

where A_o is an $n \times 1$ vector with the characteristic l-th element $a_{lo} := \frac{(d_l - n_C/n)\mu_o(w_l)}{(n_C/n)(n_U/n)}$. Since Υ_o and μ_o are unknown, $\hat{\beta}_{\Upsilon_o,\mu_o}$ is infeasible. As a feasible alternative, consider

$$\hat{\beta}_{agiv} := \left(X_U' W_U \hat{\Upsilon}_{giv}^{-1} W_U' X_U\right)^{-1} X_U' W_U \hat{\Upsilon}_{giv}^{-1} \left(\frac{n_U}{n_C} W_C' Y_C - \frac{n_U}{n} W' \hat{A}\right), \qquad \text{(2S-AGIV)}$$

where $\hat{\Upsilon}_{giv} := \hat{\Sigma}_{\hat{\beta}_{giv}} - \hat{\Psi}$, with

$$\hat{\Psi} := \frac{n^2}{n_U n_C^2} \sum_{i \in S_C} w_i (w_i' \hat{\alpha})^2 w_i' - \frac{n^2}{n_U n_C^3} (W_C' W_C \hat{\alpha})(W_C' W_C \hat{\alpha})', \qquad (3)$$

is a consistent estimator of Υ_o and, for $\hat{\alpha} = (W_C' W_C)^{-1} W_C' Y_C$, $\hat{A}$ is an $n \times 1$ column vector with the characteristic element $\hat{a}_l := \frac{w_l' \hat{\alpha}(d_l - n_C/n)}{(n_C/n)(n_U/n)}$. $w_l' \hat{\alpha}$ is a consistent estimator of $\mu_o(w_l) := E(y|w_l)$. Since $\hat{a}_l$ may be different from zero, the 2S-AGIV estimator does not belong to the GIV family.

Replacing Υ_o and μ_o with consistent estimators raises the question of how the 2S-AGIV estimator is affected. The next proposition shows that this replacement has no effect on the asymptotic precision of the 2S-AGIV estimator.

Proposition 2 (2S-AGIV - Semiparametric Efficiency). *Let Assumptions 1 to 6 hold.*

Then, $n^{1/2}(\hat{\beta}_{agiv} - \beta_o) \rightsquigarrow \mathcal{N}(0, \mathcal{I}_o^{-1})$*, where* $\mathcal{I}_o^{-1}$ *is defined in Lemma 2.*

Testing statistical hypothesis about β_o based on $\hat{\beta}_{agiv}$ requires an estimate of the variance of $\hat{\beta}_{agiv}$. One approach to obtain this estimate is to use the sample analog principle, which, from $\mathcal{I}_o^{-1}$, yields $\widehat{var}(\hat{\beta}_{agiv}) := \frac{n_U^2}{n}\left[X_U' W_U \hat{\Upsilon}_{agiv}^{-1} W_U' X_U\right]^{-1}$, where $\hat{\Upsilon}_{agiv} := \hat{\Sigma}_{\hat{\beta}_{agiv}} - \hat{\Psi}$.

The 2S-AGIV estimator is not a special case of other two-sample estimators. These include the empirical-likelihood weighted least squares estimator in Hellerstein and Imbens (1999) and the auxiliary-to-study tilting estimator in Graham et al. (2016). The empirical-likelihood weighted least squares estimator only applies to the class of models point-identifying the parameters of interest from one sample. The two-sample linear regression model does not belong to this class because it needs both samples in order to achieve point-identification. The auxiliary-to-study tilting estimator is not geared toward over-identifying models, which is the case for the two-sample linear regression model.

4. Empirical Application

This section illustrates and supports the use of the 2S-AGIV to gain insight on the relationship between income and body mass index in England. While the main purpose of the application is to demonstrate what the two-sample procedure has to offer with respect to one-sample procedures, this application is also a topic of importance for enhancing the understanding of the economic causes of obesity.

4.1 Background: The Unearned Income Effect (UIE) Hypothesis

The obesity epidemic has increasing relevance in the allocation of public resources in England. From 2014 to 2015, the annual expenditures on the treatment of obesity-related ill-health was greater than the amount spent on the police, fire service and judicial system

combined (Public Health of England, 2017). Because of their presumed significance, economic factors, in particular income, that are correlated with obesity have become a subject of research.

One possible explanation for the dependence between obesity and income comes from an individual utility maximization model of food consumption and body weight control (Lakdawalla and Philipson, 2009). In this model, the total effect of income on body weight includes the effect on food consumption and ideal weight (the unearned income effect) plus the effect of earning income on physical activity. Under the assumption that food consumption and body weight control are complements in the utility function, the unearned income effect has an inverted U-shape due to the offsetting effects on the demand for food consumption and the demand for ideal weight. For underweight individuals, growth in income shifts out food consumption, resulting in increased weight. For overweight individuals, growth in income shifts food consumption inwards reducing weight. The UIE hypothesis states that this inverted U-shaped effect dominates the effect of earning income on physical activity.

In this empirical exercise, the research question is to determine the empirical content of the UIE hypothesis. Addressing this question is relevant when assessing the potential effects of redistribution programs on average weight. Lakdawalla and Phillipson (2009) indicate that if the UIE hypothesis is correct, on one hand, redistribution in terms of unearned income -such as food stamps or cash transfers and taxes on capital gains and states- may raise the weight of individuals at both ends of the income distribution. On the other hand, redistribution in terms of earned income -such as progressive taxes on earnings- may increase the average weight of low-income individuals and decrease the weight of high-income individuals. If the UIE hypothesis is in contradiction with the data, these conjectured effects of redistribution programs on average weight will lack empirical support.

4.2 The Statistical Hypotheses and the Data

To test the inverted U-shaped relationship, consider the specification

$$bmi_i = \beta_{1o} inc_i + \beta_{2o} inc_i^2 + z_i' \gamma_o + u_i, \tag{4}$$

where bmi_i is the body mass index for individual i, inc_i represents the weekly household income, and z_i is a list of control covariates, including a constant, age, age squared, occupation and ethnicity. The disturbance term u_i is assumed to be mean-independent of income and the list of controls, as required when extending Assumption 1 to a case with control covariates. β_{1o}, β_{2o} and γ_o are the unknown coefficients. The objective is to test the null hypothesis $H_o : \beta_{2o} = 0$. If the null hypothesis is rejected in favor of the alternative $H_{A1} : \beta_{2o} < 0$, one can conclude that the data do not contradict the UIE hypothesis. If the null hypothesis is rejected in favor of the alternative $H_{A2} : \beta_{2o} > 0$, one can conclude that the data do contradict the UIE hypothesis. If the null hypothesis is not rejected, the data are not conclusive.

The Health Survey for England (HSE) is well suited to test the UIE hypothesis because it is the only annual representative survey for England that monitors the prevalence of obesity. It contains individual-level data on body mass index, income, age, occupation and ethnicity. Income, however, is reported in intervals, as in Assumption 3. During personal interviews, individuals are shown a card with 31 intervals and asked to indicate the income group they belong.[16] The analysis uses the sample for 2014 of individuals aged 20-69.

Empirical studies on obesity confronting interval-censored income, instead of (4), consider the specification $bmi_i = dinc_i' \alpha_o + z_i' \gamma_o + u_i$, where $dinc_i$ is a list of three dummy variables indicating the quartile of the income distribution to which an individual belongs (Lakdawalla and Philipson, 2009). For the sake of comparison, Table I presents ordinary least squares estimates of α_o from the HSE with robust standard errors in parentheses. Negative estimates

[16]The US National Health Interview Survey, which has been used by Lakdawalla and Philipson (2009), also reports income by intervals.

of α_{1o}, α_{3o} and non-significant estimates of α_{2o} are interpreted as evidence in favor of the UIE hypothesis. This interpretation, however, is not a statistical test.

Table I. Body Weight and Income: One-Sample Analysis

	α_{1o}	α_{3o}	α_{3o}	n_c
females	.375	-.225	-.765	2,632
	(.322)	(.265)	(.352)	
males	-.398	-1.77	-.188	2,266
	(.300)	(.355)	(.278)	

Note: Calculations based on data from the HSE 2014.

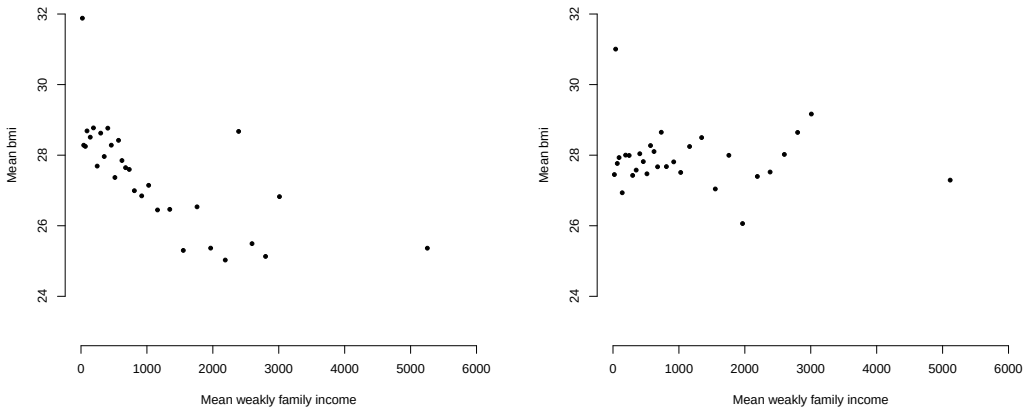
Instead of modifying the specification, this paper proposes using a second dataset with continuous measurements on income. The Family Resource Survey (FRS) is a representative sample of the population in England reporting income as a single value -as in Assumption 4. The FRS does not gather information on body mass index, but has data on age, occupation, and ethnicity. The HSE and FRS are independent samples on the same population, as in Assumption 5. Selection of an individual into one of these samples does not depend on body mass index, income, age, occupation or ethnicity. In the model presented in Section 2, there are no control covariates. These covariates are then removed from (4) by setting y_i and x_i as the residuals from projecting bmi_i and (inc_i, inc_i^2), respectively, on z_i while w_i is the vector of indicator variables for the income intervals.

In this empirical exercise, neither the existing normal parametric one-sample nor the nonparametric two-sample approaches directly apply in testing the UIE hypothesis. The existing one-sample parametric approach (see, e.g., Hsiao, 1983) would require the assumption that income and income squared are jointly -and marginaly- normally distributed, which is infeasible because neither income nor its square can take negative values. The existing two-sample nonparametric approach (e.g., Pollmann, 2015; Asher et al., 2018) would require the assumption that income has a monotonic effect on body mass index.

4.3 Empirical Results

We start the two-sample empirical analysis with a scatter plot for the means of the subsamples of female and male BMIs (calculated from the HSE) and weekly family income (calculated from the FRS), grouped by income intervals. From visual inspection, the inverted quadratic relationship between these two variables is not evident. For females, the relationship between income and body mass index seems negative, while for males, it is unclear.

Figure 1: Grouped Means - BMI and Income for Females (Left) and Males (Right)



Note: Calculations based on data from the HSE 2014 and FRS 2013/14.

Table 2 presents alternative estimates of the coefficients β_{1o} and β_{2o}. The sign of the 2SLS, 2S-GIV and 2S-AGIV estimates of β_{2o}, for both males and females, is the opposite of the sign of an inverted U-shaped relationship. The magnitude of the estimates for β_{1o} and β_{2o} is small, which is in line with evidence for the elderly in the US (see Cawley, Moran and Simon, 2010). To assess the uncertainty arising from sampling variability, notice that the 2S-AGIV is equally or more precise than the other two-sample estimators.[17] The table also presents the realized value of the asymptotic t-statistic for the null hypothesis $H_o : \beta_{2o} = 0$ (row "t-stat"). For the subsample of males, the data are inconclusive regarding the UIE

[17]Some of the moments implied by the mean-independence restriction on two of the control covariates, age and age squared, are not exploited by these estimators. It is out of the scope of this paper to exploit these extra restrictions.

hypothesis. For the subsample of females, the evidence is different. The asymptotic t-tests based on the 2S-AGIV and 2S-GIV estimators suggest that there is strong evidence against the inverted U-shaped relationship, i.e., the realized value of the t-statistic is greater than 3.090, which is the .1% critical value for the one-sided test with the alternative hypothesis $H_{A2} := \beta_{2o} > 0$. This conclusion is not available from the naive one-sample analysis based on the OLS estimator.

Table II. Body Weight and Income: Two-Sample Analysis

	2SLS		2S-GIV		2S-AGIV	
	females	males	females	males	females	males
β_{1o}	-1.6e-03	-1.9e-04	-2.2e-03	-3.2e-04	-2.2e-03	-3.2e-04
	(4.7e-04)	(3.7e-04)	(4.2e-04)	(3.1e-04)	(4.2e-04)	(3.1e-04)
β_{2o}	1.4e-07	3.4e-10	2.1e-07	2.6e-08	2.1e-07	2.9e-08
	(5.9e-08)	(5.4e-08)	(5.5e-08)	(4.8e-08)	(5.5e-08)	(4.8e-08)
t-stat	2.43	.006	3.84	.538	3.84	.600
n_C	2,632	2,266	2,632	2,266	2,632	2,266
n_U	11,388	10,142	11,388	10,142	11,388	10,142

Note: Calculations based on data from the HSE 2014 and FRS 2013/14.

The result for the subsample of females illustrates the main point of this paper, namely, how using a second sample enhances the one-sample analysis.

5. Simulation Study

To verify the theoretical properties of the 2S-AGIV estimator, and to evaluate its numerical performance, this section reports the results of Monte Carlo experiments.

5.1 Design of Experiments

The underlying model used in all of the experiments is given by Assumptions 1 to 6, with u given x_l distributed as $\mathcal{N}(0, \sigma_l^2)$ and $\sigma_l^2 := a_o \times (.1 + x_l^2)^{c_o}$, where a_o and c_o are constants such that u is distributed as $\mathcal{N}(0, 5)$. The parameters varying across experiments are the

sample sizes n_C and n_U, the distribution of (x, w), the coefficient β_o, and the constants a_o and c_o. The total sample size $n = n_C + n_U$ varies between $n = 1,000$ and $n = 4,000$. When $n_C \neq n_U$, the theory predicts that the 2S-GIV efficiency loss is amplified. When $n_C = n_U$, the theory predicts that efficiency loss is minimized. The covariate follows either a normal distribution $\mathcal{N}(0, 2)$ with a zero mean and a variance of 2 or a Student $\mathcal{T}_4$ distribution with 4 degrees of freedom (with a zero mean and a variance of 2). The interval-censoring scheme has the following $B + 1 = 6$ categories:

$$g_L(x) = \begin{cases} L_1 = -\infty & \text{if } -\infty < x < -1 \\ L_2 = -1 & \text{if } -1 < x < -.5 \\ L_3 = -.5 & \text{if } -.5 < x < 0 \\ L_4 = 0 & \text{if } 0 < x < .5 \\ L_5 = .5 & \text{if } .5 < x < 1 \\ L_6 = 1 & \text{if } 1 < x < \infty \end{cases} ; g_U(x) = \begin{cases} U_1 = -1 & \text{if } -\infty < x < -1 \\ U_2 = -.5 & \text{if } -1 < x < -.5 \\ U_3 = 0 & \text{if } -.5 < x < 0 \\ U_4 = .5 & \text{if } 0 < x < .5 \\ U_5 = 1 & \text{if } .5 < x < 1 \\ U_6 = \infty & \text{if } 1 < x < \infty \end{cases}$$

β_o varies between $\beta_o = 1$ (the theory predicts that the efficiency loss is amplified) and $\beta_o = 0$ (efficiency loss is minimized). The constants a_o and c_o control for the degree of conditional heteroscedasticity: $c_o = 0$ corresponds to conditional homoscedasticity, and $c_o = 1$ corresponds to conditional heteroscedasticity. The number of Monte Carlo replications is 20,000.

Open-ended intervals (such as $L_1 = \infty$ or $U_6 = \infty$), non-normal covariates, and heteroscedasticity are common phenomena in applications using survey data on income; thus, these elements of the design are particularly relevant to practice. $B = 5$ is less than the number of categories commonly found in survey data, which also makes the experiments relevant to practice.

5.2 Simulation Results

Table III reports the bias, standard deviation (sd) and the root mean squared error (rmse) of the estimators. It also reports the coverage and average length of the associated confidence intervals, and the skewness and excess kurtosis of the standardized sampling errors. The results are in line with the theoretical predictions. The sd of the 2S-AGIV estimator is lower than that of the other two estimators (see column (2)), except in the homoscedastic case (see Experiments 7-9). The bias of the 2S-AGIV estimator is lower than that the 2S-GIV but may be higher than that of the 2SLS estimator (see column (1)). Overall, the 2S-AGIV offers gains in rmse except in the homoscedastic experiments (see column (3)). The coverage level of the 2S-AGIV confidence interval is close to the nominal 95% level (see column 4), with some undercoverage for the smallest sample size (see experiment 11). This undercoverage may be explained by the imprecision in the estimation of Υ_o. As predicted from the theory, the average length of the 2S-AGIV confidence interval is smaller than that of the other confidence intervals (see column 5). The skewness and excess kurtosis of the sampling errors are nonzero (see columns (6) and (7)), which suggests that there is room to refine the normal approximation in Proposition 2. However, these refinements are beyond the scope of this paper.

Table III. Summary of Simulation Results - Estimators

	(1) bias	(2) sd	(3) rmse	(4) cove	(5) length	(6) skew	(7) kur
Experiment 1: $n = 4,000$, $\kappa_o = .2$, *x is Normal, u|x is Heteroscedastic,* $\beta_0 = 1$.							
2SLS	-.002	.208	.208	95%	.816	.023	.042
2S-GIV	-.017	.076	.078	93%	.290	.037	.003
2S-AGIV	-.003	.066	.066	93%	.244	.037	.013
Experiment 2: $n = 4,000$, $\kappa_o = .4$, *x is Normal, u|x is Heteroscedastic,* $\beta_0 = 1$.							
2SLS	-.001	.148	.148	95%	.582	.063	.003
2S-GIV	-.007	.057	.057	94%	.219	.063	.001
2S-AGIV	.001	.047	.047	94%	.179	.027	-.007
Experiment 3: $n = 4,000$, $\kappa_o = .5$, *x is Normal, u|x is Heteroscedastic,*$\beta_0 = 1$.							
2SLS	.000	.132	.132	95%	.523	.070	.018
2S-GIV	-.005	.053	.053	94%	.204	.067	.048
2S-AGIV	.003	.043	.043	94%	.164	.029	.002
Experiment 4: $n = 4,000$, $\kappa_o = .2$, *x is Student, u|x is Heteroscedastic,*$\beta_0 = 1$.							
2SLS	-.014	.964	.964	96%	2.66	-3.32	163
2S-GIV	-.014	.100	.101	94 %	.382	.033	-.040
2S-AGIV	-.005	.097	.097	94%	.364	.020	-.051
Experiment 5: $n = 4,000$, $\kappa_o = .4$, *x is Student, u|x is Heteroscedastic,*$\beta_0 = 1$.							
2SLS	-.010	.827	.827	96%	1.94	-25.2	1,997
2S-GIV	-.007	.071	.071	94%	.271	.026	-.015
2S-AGIV	-.002	.067	.067	94%	.255	.008	-.016
Experiment 6: $n = 4,000$, $\kappa_o = .5$, *x is Student, u|x is Heteroscedastic,*$\beta_0 = 1$.							
2SLS	-.010	.714	.714	96%	1.77	-19.9	1,428
2S-GIV	-.006	.064	.064	94%	.249	.062	.044
2S-AGIV	-.001	.061	.061	94%	.232	.026	.015
Experiment 7: $n = 4,000$, $\kappa_o = .2$, *x is Normal, u|x is Homoscedastic,* $\beta_0 = 1$.							
2SLS	-.002	.099	.099	96%	.405	.010	-.036
2S-GIV	-.003	.104	.104	95%	.403	.022	-.011
2S-AGIV	-.002	.100	.100	95%	.387	.012	-.024
Experiment 8: $n = 4,000$, $\kappa_o = .4$, *x is Normal, u|x is Homoscedastic,* $\beta_0 = 1$.							
2SLS	.000	.071	.071	96%	.405	.010	-.036
2S-GIV	-.001	.075	.075	95%	.403	.022	-.011
2S-AGIV	-.001	.071	.071	95%	.387	.012	-.024
Experiment 9: $n = 4,000$, $\kappa_o = .5$ *x is Normal, u|x is Homoscedastic,* $\beta_0 = 1$.							
2SLS	.000	.064	.064	96%	.265	.027	-.024
2S-GIV	-.001	.068	.068	95%	.264	.058	-.044
2S-AGIV	-.001	.064	.064	95%	.249	.031	-.021
Experiment 10: $n = 4,000$, $\kappa_o = .2$, *x is Normal, u|x is Heteroscedastic,* $\beta_0 = 0$.							
2SLS	-.002	.207	.207	95%	.804	-.002	.043
2S-GIV	.000	.057	.057	95%	.220	-.004	.048
2S-AGIV	.000	.059	.059	93%	.219	-.004	.042
Experiment 11: $n = 4,000$, $\kappa_o = .05$, *x is Normal, u|x is Heteroscedastic,* $\beta_0 = 1$.							
2SLS	.001	.414	.414	95%	1.60	.046	.143
2S-GIV	-.074	.155	.172	85%	.526	-.025	.177
2S-AGIV	-.024	.142	.144	86%	.436	.001	.357

Note: This table reports the bias, standard deviation (sd), root mean squared error (rmse) of the estimators in the text and the skewness (skew) and excess kurtosis (kur) of their standardized sampling errors. It also reports the coverage (cove) and average length of the 95%-level confidence intervals.

6. Summary and Directions for Further Research

In a linear regression model with interval-censored covariates, information from an auxiliary uncensored sample (not necessarely measuring the outcome variable) can be used to mitigate the undesirable effects of interval-censoring. This paper uses this observation to construct a consistent, asymptotically normal and semiparametrically efficient two-sample instrumental variable estimator. This estimator is a semiparametric alternative to existing parametric point-identifying (see, e.g. Hsiao, 1983) procedures. An application shows that the new two-sample estimator can reject an economic hypothesis of interest in a context where existing procedures do not apply.

The following refinements illustrate both the potential and the limitations of the main idea of this paper, namely, the combination of samples to restore point-identification in the presence of interval-censored covariates. The identification result in Section 2 can be generalized to the class of moment models

$$E[m_C(y;\theta_o) - m_U(x,z;\theta_o)|f(x,z)] = 0, \tag{MM}$$

where $m_C(\cdot)$ and $m_U(\cdot)$ are known functions, up to the finite-dimensional parameter θ_o, z is a list of variables observed in both samples, and $f(x,z)$ is a known function. The class MM includes the linear regression model, i.e., $\theta_o = (\beta_o, \gamma_o)$, $m_C(y;\theta_o) = y$, $m_U(x,z;\theta_o) = x'\beta_o + z'\gamma_o$, the logit model, i.e., $m_C(y;\theta_o) = y$ and $m_U(x,z;\theta_o) = exp(x'\beta_o + z'\gamma_o)/[1 + exp(x'\beta_o + z'\gamma_o)]$ where y is a binary random variable, the Weibull mixed proportional hazard model, i.e., $m_C(y;\theta_o) = \tau_o y^{\alpha_o}$ and $m_U(x,z;\theta_o) = x'\beta_o + z'\gamma_o$ where y is a positive random variable, and the instrumental variable model with covariates y_2 are only observed in the censored sample, i.e., $m_C(y;\theta_o) = y_1 - y_2\gamma_o$ and $m_U(x,z;\theta_o) = x'\beta_o$ where $y = (y_1, y_2)$. MM excludes quantile regression models. By paralleling the identification result in Proposition 1, the linear model can still point identify the coefficient of interest if x and z are not jointly

observed in the uncensored sample after changing $m_C(y,\theta) = y$ to $m_C(y,z,\theta) = y - z'\gamma_o$.

When $f(x,z) = (x,z)$ and z is discrete, MM is equivalent to a finite list of feasible unconditional moment restrictions, in which one replaces z by a list of dummy variables coding its values.[18] The efficiency bound for θ_o can still be obtained after specializing the result in Graham et al. (2016, Theorem 1) to the problem of interval-censored covariates. When z is continuous, MM is equivalent to an infinite list of feasible unconditional moment restrictions. Constructing an efficient estimator in this case requires more elaboration because one needs to rely on a user-chosen parameter to select a finite list of moments. The analysis of this case is left for future research.

[18] A feasible moment restriction is one with a feasible sample analog.

References

Asher, S., P. Novosad and C. Rafkin (2018): "Partial Identification of Expectations with Interval Data", *manuscript*. Retrived from https://arxiv.org/pdf/1802.10490.pdf

Beresteanu, A., I. Molchanov, and F. Molinari (2011): "Sharp Identification Regions in Models with Convex Moment Predictions", *Econometrica, 79*, 1785-1821.

Bhattacharya, D and Y.-Y. Lee (2018): "Applied Welfare Analysis for Discrete Choice with Interval-Data on Income", *manuscript*. Retrived from
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3167071

Bontemps, C., T. Magnac and E. Maurin (2012): "Set Identified Linear Models", *Econometrica, 80*, 1129-1155.

Carroll, R, D. Ruppert, L. Stefanski, and M. Crainiceanu (2006): *Measurement Error in Nonlinear Models: A Modern Perspective* (2nd ed.). Boca Raton, FL: CRC.

Cawley, J., J. Moran and K. Simon (2010): "The Impact of Income on the Weight of Elderly Americans", *Health Economics, 19*, 979-993.

Cerquera, D. F. Laisney and H. Ullrich (2015): "A Note on a Regression with Interval Data on a Regressor", *manuscript*. Retrived from
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2473768

Chaudhuri, S. and D. Guilkey (2016): "GMM with Multiple Missing Variables", *Journal of Applied Econometrics, 31*, 678-706.

Chen, X. H. Hong and A. Tarozzi (2008): "Semiparametric Efficiency in GMM models with Auxiliary Data", *Annals of Statistics, 36*, 808-843.

Choi, J., J. Gu and S. Shen (2017): "Weak-Instruments Robust Inference for Two-Sample Instrumental Variable Regression", *Journal of Applied Econometrics, 33*, 109-105.

Dardanoni, V., G. de Luca, S. Modica and F. Peracchi (2015): "Model averaging estimation of generalized linear models with imputed covariates", *Journal of Econometrics, 184*, 452-

463.

Devereux, P. and G. Tripathi (2009): "Optimally Combining Censored and Uncensored Datasets", *Journal of Econometrics, 151*, 17-32.

Graham, B.(2011): "Efficiency Bounds for Missing Data Models with Semiparametric Restrictions", *Econometrica, 79*, 437-452.

Graham, B., C. Pinto, and D. Egel (2011): "Efficient Estimation of Data Combination Models by the Method of Auxiliary-to-Study Tilting", *NBER Working Paper Series*, Working Paper 16928.

Graham, B., C. Pinto, and D. Egel (2012): "Inverse Probability Tilting for Moment Conditions Models with Missing Data", *Review of Economic Studies, 79*, 1053-1079.

Graham, B., C. Pinto, and D. Egel (2016): "Efficient Estimation of Data Combination Models by the Method of Auxiliary-to-Study Tilting", *Journal of Business and Economic Statistics, 34*, 288-301.

Hellerstein, J. and G. Imbens (1999): "Imposing Moment Restrictions from Auxiliary Data by Weighting", *The Review of Economics and Statistics, 81*, 1-14.

Hsiao, C. (1983): "Regression Analysis with a Categorized Explanatory Variable" in S. Karlin, T. Ameniya and L. Goodman (Eds.): *Studies in Econometrics, Time Series and Multivariate Statistics*. San Diego, CA: Academic Press.

Inoue, A. and G. Solon (2010): "Two-Sample Instrumental Variable Estimators", *The Review of Economics and Statistics, 92* 557-561.

Juster, T. and J. Smith (1997): "Improving the Quality of Economic Data: Lessons from HRS and AHEAD", *Journal of the American Statistical Association, 92*, 1268-1278.

Kaido, H. (2017): "Asymptotically Efficient Estimation of Weighted Average Derivatives with an Interval Censored Variable", *Econometric Theory, 33*, 1218-1241.

Klevmarken, N. (1982): "Missing Variables and Two-Stages Least Squares Estimation from More than One Dataset", Working Paper No. 62, Research Institute of Industrial Economics,

Stockholm.

Lakdawalla, D. and T. Philipson (2009): "The Growth of Obesity and Technological Change", *Economics and Human Biology, 7*, 283-293.

Lepanjuuri, K. P. Cornick, C. Byron, I. Templeton and J. Hurn (2016): *National Travel Survey 2015 Technical Report*, UK Department for Transport.

Magnac, T. and E. Maurin (2008): "Partial Identification in Binary Models: Discrete Regressors and Interval Data", *Review of Economic Studies, 75*, 835-864.

Manski, C. and E. Tamer (2002): "Inference on Regressions with Interval Data on a Regressor or Outcome", *Econometrica, 70*, 519-546.

Pacini, D. and F. Windmeijer (2016): "Robust Inference for the Two-Sample 2SLS Estimator", *Economic Letters, 146*, 50-54.

Pollmann, D. (2015): "Identification and Estimation with an Interval-censored Regressor when its Marginal Distribution is Known", *manuscript*.

Public Health of England (2017): "Health Matters: Obesity and the Food Environment", Retrived from https://www.gov.uk/government/publications/health-matters-obesity-and-the-food-environment/health-matters-obesity-and-the-food-environment–2.

Ridder, G. and R. Moffitt (2007): "The Econometrics of Data Combination" in Heckman, J. and E. Leamer (eds.), *Handbook of Econometrics, Volume 6B*. Amsterdam: North-Holland.

Rigobon, R. and T. Stoker (2009): "Bias from Censored Regressors", *Journal of Business and Economic Statistics, 27*, 340-353.

Robins, J., A. Rotnitzky and L. Zhao (1994): "Estimation of Regression Coefficients when Some Regressors are not Always Observed", *Journal of the American Statistical Association, 89* 846-866.

Winter, J. (2002): "Bracketing Effects in Categorized Survey Questions and the Measurement of Economic Quantities", *manuscript*.

Appendix A: Proof of Propositions

Proof of Proposition 1 (Point Identification). Since w is a measurable transformation of x, from $E(u|x) = 0$ in Assumption 1 and after replacing $u = y - x'\beta_o$, one has $E[y - x'\beta_o|w] = 0$. Furthermore, since d is independent of (y, x) (see Assumption 5), one has

$$E\left[\frac{d}{\kappa_o}y - \frac{(1-d)}{(1-\kappa_o)}x'\beta_o\middle| w\right] = 0$$

Since w is a vector of indicator variables, the conditional moment restriction in the latter display holds if and only if

$$E\left[\frac{d}{\kappa_o}wy\right] - E\left[\frac{(1-d)}{(1-\kappa_o)}wx'\right]\beta_o = 0. \tag{WMR}$$

Under Assumption 2 and 5, the censored sample in Assumption 3 is identifying $E\left[\frac{d}{\kappa_o}wy\right]$ and the uncensored sample in Assumption 4 is identifying $E\left[\frac{(1-d)}{(1-\kappa_o)}wx'\right]$. Hence, the model is point-identifying β_o if $E[w(y - x'\beta)] \neq 0$ for all $\beta \neq \beta_o$ because $0 < \kappa_o < 1$. One can rewrite this condition as $E\big[w\big(y - x'(\beta - \beta_o)\big)\big] \neq 0$ for all $\beta - \beta_o \neq 0$. For this condition to hold, it is necessary and sufficient that $E(xw')E(ww')^{-1}E(wx')$ is full rank. This must be the case if $E(ww')$ and $E(wx')$ are full rank, which indeed are under Assumption 6. □

Proof of Proposition 2 (2S-AGIV- Asymptotic Efficiency). The 2S-AGIV estimator satisfies

$$0 = n_U^{-1}\sum_{j\in S_U} x_j w_j' \hat{\Upsilon}^{-1}\left(n_C^{-1}\sum_{i\in S_C} w_i y_i - n^{-1}\sum_{l\in S} w_l \hat{a}_l - n_U^{-1}\sum_{j\in S_U} w_j x_j'\hat{\beta}_{agiv}\right)$$

Add-and-subtract $n_U^{-1}\sum_{j\in S_U} w_j x_j \beta_o$ inside the parenthesis to obtain:

$$0 = n_U^{-1}\sum_{j\in S_U} x_j w_j' \hat{\Upsilon}^{-1}\left(n_C^{-1}\sum_{i\in S_C} w_i y_i - n_U^{-1}\sum_{j\in S_U} w_j x_j'\beta_o - n^{-1}\sum_{l\in S} w_l \hat{a}_l - n_U^{-1}\sum_{j\in S_U} w_j x_j'(\hat{\beta}_{agiv} - \beta_o)\right).$$

Working out $(\hat{\beta}_{agiv} - \beta_o)$ and multiplying both sides by $n^{1/2}$ yield:

$$n^{1/2}(\hat{\beta}_{agiv} - \beta_o) = \left[n_U^{-1} \sum_{j \in S_U} x_j w_j' \hat{\Upsilon}^{-1} n_U^{-1} \sum_{j \in S_U} w_j x_j' \right]^{-1} n_U^{-1} \sum_{\in S_U} x_j w_j' \hat{\Upsilon}^{-1} \qquad \text{(First Term)}$$

$$\left(\frac{n^{1/2}}{n_C^{1/2}} n_C^{1/2} n_C^{-1} \sum_{i \in S_C} w_i y_i - \frac{n^{1/2}}{n_U^{1/2}} n_U^{1/2} n_U^{-1} \sum_{j \in S_U} w_j x_j' \beta_o - \frac{n^{1/2}}{n} \sum_{l \in S} w_l \hat{a}_l \right). \qquad \text{(Second Term)}$$

By Assumptions 1, 3 and 4, an application of the Law of Large Numbers and the Continuous Mapping Theorem to the First Term yields

$$\text{First Term} \rightarrow_P \left[E(xw') \Upsilon_o^{-1} E(wx') \right]^{-1} E(xw') \Upsilon_o^{-1}$$

To derive the asymptotic variance, it suffices to find the asymptotic distribution of the Second Term. Expanding around a_{lo} gives:

$$\frac{n^{1/2}}{n} \sum_{l \in S} w_l \hat{a}_l = \frac{n^{1/2}}{n} \sum_{l \in S} w_l a_{lo} + n^{-1} \sum_{l \in S} w_l \frac{(d - \kappa_o)}{\kappa_o (1 - \kappa_o)} w_l' n^{1/2} (\hat{\alpha} - \alpha_o) + o_P(1)$$

The sum $n^{-1} \sum_{l \in S} w_l \frac{(d - \kappa_o)}{\kappa_o (1 - \kappa_o)} w_l'$ in the right hand side converges in probability to $E\left(w_l \frac{(d - \kappa_o)}{\kappa_o (1 - \kappa_o)} w_l' \right)$. Since d and w are independent and $\kappa_o := E(d)$ -see Assumption 5 (a)-, one has $E\left(w_l \frac{(d - \kappa_o)}{\kappa_o (1 - \kappa_o)} w_l' \right) = E(ww') E(d - \kappa) / V(d) = 0$. Hence, $n^{-1} \sum_{l \in S} w_l \frac{(d - \kappa_o)}{\kappa_o (1 - \kappa_o)} w_l' = o_P(1)$. Moreover, since the uncensored sample is iid with finite second moments -see Assumptions 1(c) and 4-, the OLS estimator $\hat{\alpha}$ is bounded in probability: $n^{1/2}(\hat{\alpha} - \alpha_o) = O_P(1)$. One then can write

$$\frac{n^{1/2}}{n} \sum_{l \in S} w_l \hat{a}_l = \frac{n^{1/2}}{n} \sum_{l \in S} w_l a_{lo} + o_P(1) O_P(1) + o_P(1) = \frac{n^{1/2}}{n} \sum_{l \in S} w_l a_{lo} + o_P(1)$$

The Second Term is then asymptotically equivalent to:

$$\left(\frac{n^{1/2}}{n_C^{1/2}} n_C^{1/2} n_C^{-1} \sum_{i \in S_C} w_i y_i - \frac{n^{1/2}}{n_U^{1/2}} n_U^{1/2} n_U^{-1} \sum_{j \in S_U} w_j x_j' \beta_o - \frac{n^{1/2}}{n} \sum_{l \in S} w_l a_{lo}\right),$$

which is the square root of the sample size multiplied by the sample analog of the augmented weighted moment restriction. An application of the Central Limit Theorem yields:

$$\text{Second Term} \rightsquigarrow \mathcal{N}(0, \Upsilon_o),$$

where the variance has been derived in Lemma 3.

Using the Slutzky Lemma to combine the limits in probability and in distribution for the First and Second Terms yields:

$$\begin{aligned} n^{1/2}(\hat{\beta}_{agiv} - \beta_o) &\rightsquigarrow \mathcal{N}\big(0, \big[E(xw')\Upsilon_o^{-1}E(wx')\big]^{-1} E(xw')\Upsilon_o^{-1}\Upsilon_o\Upsilon_o^{-1}E(wx')\big[E(xw')\Upsilon_o^{-1}E(wx')\big]^{-1}\big) \\ &\rightsquigarrow \mathcal{N}\big(0, \big[E(xw')\Upsilon_o^{-1}E(wx')\big]^{-1}\big). \end{aligned}$$

□

Appendix B: Misspecification Testing

The use of the 2S-AGIV estimator when the two-sample linear regression model is actually incorrect leads to invalid inferences about β_o. This suggests that the two-sample linear regression model should be tested. Building upon the well-known Hansen and Hausman misspecification tests, there are at least two procedures for testing the invalidity of the two-sample linear regression model.

In the first procedure, the null hypothesis is $E[dwy/\kappa_o - (1-d)x\beta_o/(1-\kappa_o)] = 0$. The procedure is based on the overidentification statistic:

$$J := n \times \left(\frac{W_C'Y_C}{n_C} - \frac{W_U'X_U\hat{\beta}_{agiv}}{n_U} - \frac{W'\hat{A}}{n}\right)' \hat{\Upsilon}_{agiv}^{-1} \left(\frac{W_C'Y_C}{n_C} - \frac{W_U'X_U\hat{\beta}_{agiv}}{n_U} - \frac{W'\hat{A}}{n}\right).$$

Under Assumptions 1 to 6, J converges in distribution to a chi-squared random variable (denoted $J \rightsquigarrow \chi_{df}$) with degrees of freedom equal to the rank of $\hat{\Upsilon}_{agiv}^{-1}$ (i.e, $df = B - 1$). For a pre-specified level a, this suggests the asymptotic $100 \times a\%$ misspecification test $T_J = 1(\chi_{B-1}^{-1}(1-a) \leq J)$.

The second procedure is based on the Hausman-type statistic:

$$H := (\hat{\beta}_{agiv} - \hat{\beta}_{2sls})'[\widehat{var}(\hat{\beta}_{2sls}) - \widehat{var}(\hat{\beta}_{agiv})]^{-1}(\hat{\beta}_{agiv} - \hat{\beta}_{2sls}).$$

where $\widehat{var}(\hat{\beta}_{2sls})$ is a consistent estimator of the variance of the 2SLS estimator. The null hypothesis in this case is that the limit in probability of $\hat{\beta}_{agiv}$ and $\hat{\beta}_{2sls}$ coincide. Under correct specification, H converges in distribution to a chi-squared random variable with degrees of freedom equal to the rank of $avar(\hat{\beta}_{2sls}) - \Upsilon_o^{-1}$.[19] Under misspecification, the 2S-AGIV and 2SLS estimators converge in probability to different values, so that the distance between $\hat{\beta}_{2sls}$ and $\hat{\beta}_{agiv}$ is nonzero in large samples. This suggests the asymptotic $100 \times a\%$

[19] Convergence, however, is not uniform, which may create size distortions, in particular, at $\beta_o = 0$.

misspecification test $T_H = 1(\chi_1^{-1}(1-a) \leq H)$.

Rejection by either of these tests, i.e., $T_J = 1$ or $T_H = 1$, indicates that the linear regression model is invalid or that w_i and w_j are not identically distributed in the censored and uncensored samples. When the misspecification tests do not reject, i.e., $T_J = 0$ and $T_H = 0$, one may decide to proceed with the calculation and interpretation of the confidence interval:

$$\left[\hat{\beta}_{agiv} \pm 1.96\sqrt{\widehat{var}\left(\hat{\beta}_{agiv}\right)}\right]. \qquad \text{(G-Interval)}$$

The two-stage G-interval, which uses a misspecification test in the first step and the G-interval in the second step, suffers from a coverage distortion, which means that its actual coverage probability may be lower than the pre-specified confidence level. One way of overcoming this coverage distortion is by inverting the two-sample continuous updating objective function:

$$S(\beta) := n \times \left(\frac{W_C' Y_C}{n_C} - \frac{W_U' X_U \beta}{n_U} - \frac{W'\hat{A}}{n}\right)' (\hat{\Sigma}_\beta - \hat{\Psi})^{-1} \left(\frac{W_C' Y_C}{n_C} - \frac{W_U' X_U \beta}{n_U} - \frac{W'\hat{A}}{n}\right),$$

When $\beta = \beta_o$, $S(\beta)$ converges in distribution to a chi-squared distribution with degrees of freedom equal to the rank of Υ_o ($S(\beta_o) \rightsquigarrow \chi_B$). Hence,

$$\text{S-set}_a := \{\beta \in \mathbb{R} : S(\beta) \leq \chi_B^{-1}(1-a)\} \qquad \text{(One-Step S-set)}$$

is an asymptotically valid $100 \times (1-a)\%$ confidence set. The S-set consists of coefficients values at which data fails to reject the model and the null hypothesis $\beta_o = \beta$. If the model is misspecified, the S-set can be empty. If the model is under-identifying β_o, the S-set can be the real line. The case of a non-empty small S-set requires care in interpretation. The S-set could be small either because the model is correctly specified and β_o is precisely estimated

or because the model is misspecified but the data are not informative to reject the model.